Research Paper | Engineering | E-ISSN No : 2454-9916 | Volume : 7 | Issue : 5 | May 2021

PREDICTION OF DIABETES USING MACHINE LEARNING

Anamika[1], Isha Tyagi[1]

[1]Amity School of Engineering Noida, Uttar Pradesh, India.

ABSTRACT

Diabetes is observed to be among the most perilous diseases and chronic diseases. A multitude of problems arise if the disease is left unattended and untreated. The prosaic task of identification of the problem aggregates to a patient visiting a doctor in a medical center for deliberation. However, the rise of machine learning methodologies solve this severe problem. The incentive of this study is to examine the model which can prefigure the plausibility of diabetes in patients with maximal accuracy. Thereupon, the four algorithms namely Decision Tree, Random Forest, Naive Bayes and Adaboost Classifier are utilized in this research for predicting diabetes at an initial stage. This paper aims at testing befitting algorithms for the prediction of diabetes.

Experiments are conducted on two datasets namely Pima India Diabetes Database (PIDD) which is referenced from UCI machine learning repository and an auxiliary database. The efficiency of all four algorithms are assessed on the basis of Accuracy, Precision, Recall and F-Measure. Using these all four algorithms discussed above the result that was acquired reveal Adaboost Classifier exceeds the other algorithms with the highest accuracy of 85.5% for PIDD and 95.4% for the aforementioned dataset. The result that was obtained using Receiver Operating Characteristic (ROC) curves in a sequential manner.

KEYWORDS: Diabetes; Random Forest; Naive Bayes; Decision Tree; Adaboost; Accuracy.

1. INTRODUCTION:

Diabetes, scientifically known as Diabetes Mellitus is a recurrent chronic disease which threatens human health. It is a condition that impedes body's ability to process blood sugar. Malfunction of insulin hormone causes increase in blood sugar levels. Unprocessed high blood glucose from diabetes have dire effects and cause disfunctioning of various nerves, especially eyes, kidneys, and other organs (Krasteva et al., 2011).With the development and progression of living standards , diabetes has become intensively common among people. Obesity, absence of physical activity, smoking, unhealthy diet integrates towards diabetic conditions leading to complications in many parts of the body and amplifying the risk of premature death. According to the Hindu, approximated number of adult people suffering from diabetes in India are evaluated at 77 million. The frequency in urban areas is in the range of 10.9%-14.2%, and rural areas in range of 3.0-7.8%.

Diabetes is categorised in two categories, type 1 diabetes (T1D) and type 2 diabetes (T2D) . Type 1 diabetes, otherwise called juvenile diabetes, is a chronic situation which occurs when pancreas fail to produce insulin causing the patients to be insulin-dependent (Iancu et al., 2008).The symptoms include increased thirst and hunger, frequent urination and blurry vision. Type 2 diabetes arises from the lifestyle factors and genetics where the body doesn't utilize insulin correctly, developing abnormal blood sugar levels. It occurs more prominently in individuals aged 45 or above (Robertson et al., 2011).

Thereupon, means to rapidly discover and analyze the condition of diabetes and risks involved with it is a topic worthy of study. Alarming rise in the cases of diabetes globally, has made it important to come up with solutions for early stage detection of the disease. Machine learning has been very helpful in the prediction of a lot of diseases with the help of its analysis tools it has become a boon in the medical field. With the aid of various machine learning algorithms its usage can be extended for the prediction of the Diabetes Mellitus and thus, help in curbing its rapidity.

Numerous researchers have applied variegated algorithms for prediction of diabetes mellitus, using algorithms viz SVM, Decision tree, Naïve Bayes, Decision forest, PCA, J48etc. Mujumdar & Vaidehi (2019, p. 293) juxtaposed between machine learning algorithms for predicting presence of diabetes. Zou et al. (2018, p. 515) recognized diabetes from normal people by using principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) for dimensionality reduction. Joshi & Chawan, (2018, p. 12) used three different supervised machine learning methods namely SVM, Logistic regression, ANN.

Machine learning algorithms like decision tree has become very popular in the prediction of Diabetes Mellitus, due to its apt classification. This research deals with gestation diabetes and makes use of data of females above the age of 21 years.

The algorithms used in this research include Decision tree, Random forest, Naïve Bayes and AdaBoost. Experimental efficiency of the above mentioned algorithms are compared and a conclusion is drawn.

The first dataset used for this study contains medical detail of 768 instances which are female patients. The dataset contains 8 attributes where value of one class '0' treated as tested negative for diabetes and value of another class '1' is treated as tested positive for diabetes.

The second dataset used for this study contains 15000 observations of females above the age of 21. It has 9 attributes i.e PatientID, Pregnancies, PlasmaGlucose, DiastolicBloodPressure, TricepsThickness, SerumInsulin, BMI, DiabetesPedigree and Age.

2. RELATED WORK:

To address the problem of growing cases of diabetes a large body of research had been conducted for its detection. Most prior work in this area focused on using horde of learning techniques to obtain human relevant judgments for evaluation of diabetes.

Many researchers are using varied machine learning techniques and algorithms to obtain relevant judgments and results for evaluation of presence of diabetes .Several approaches provided support for arriving at optimized conclusions.

For example, working on designing a model for the prediction of Diabetes Mellitus where the research aimed at establishing a relation between age and diabetes. It made use of Decision Tree for the prediction of diabetes which gives satisfactory results. Regression technique was added with a randomization code that helped in the prediction of age along with the prediction of diabetes Orabi et al. in (2009).

Q Zhou et al. designed prediction model for Diabetes Mellitus using the dataset from hospital physical examination in Luzhou, China. Decision tree , random forest and neural network were used for the prediction .

D Sisodia et al. in (2018) researched on a model for the early prediction of diabetes using SVM, Naïve Bayes and Decision Tree on Prima Indian dataset. The research suggested that Naïve Bayes had the highest prediction among the afore mentioned algorithms. Joshi & Chawan,(2018, p. 12) centralized their study essentially on three supervised machine learning methods: SVM, Logistic regression, ANN.

N Yuvraj et al. In(2017) worked on the prediction of diabetes using Hadoop clustering to be able to deal with enormous amount of unstructured data and find the applicability in modern healthcare systems. Pima India Diabetes Dataset was used for the prediction of diabetes in this model.

Multiple algorithms were applied on the dataset and efficiency of each algorithm was calculated on the basis of accuracy, precision, recall, F Measure and Receiver Operating Curve. In accordance to research done globally, algorithms like Naïve Bayes, Random Forest , Decision Tree and AdaBoost were applied to draw inferences that maybe most accommodating in prediction with meticulousness.

Copyright© 2021, IERJ. This open-access article is published under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License which permits Share (copy and redistribute the material in any medium or format) and Adapt (remix, transform, and build upon the material) under the Attribution-NonCommercial terms.

The work exhibited in this paper, takes reference of previous research to provide adequate prediction based on how availability of information impacts the study conducted showcasing the impact on the decisions of the result. Further, it aims to expand the amplitude of diabetes prediction and deliver the results with exactitude.

3. METHODOLOGY USED:

3.1 Model Diagram:

The flow of the diagram in the figure showcases the analysis performed in evaluating the model.

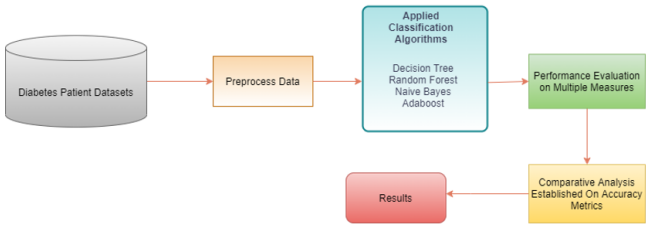

FIG.1. Proffered procedure is encapsulated in the form of a model diagram.

3.2 Brief Explanation of Algorithms Used :

3.2.1 Decision Tree:

Decision tree is part of Supervised learning algorithm. It is largely used for the classification problems. It is a type of shape of a tree which has a node, known as a leaf or a decision node. Its purpose is to separate the population into two or more similar sets based on the most notable predictors by calculating the entropy of each and every attributes. Further, the dataset is distributed by means of the variables with the most prominent information gain or minimum entropy. The two steps mentioned, are done recursively with the remaining attributes .

It uses two nodes for classification: internal and external nodes that is linked to one another. The inside or internal nodes represent decisions. It implies to be the decision making part and the leaf nodes are associated with the labels.

The assessed performance of Decision tree classifier is depicted by the confusion matrix below:

Table 1. Confusion Matrix of Decision Tree

	TP	FP	TN	FN
DATASET 1	163	27	54	29
DATASET 2	3253	290	1379	403

3.2.2 Random Forest:

Random Forest is one of the ensemble boosting classification algorithms. It works by selecting a training subset randomly from the specified dataset. It sequentially trains the AdaBoost learning model by designating the training set formulated on the accurate prediction of the prior training.

It develops a grove of decision trees and predicts the outcome which is centered on the decision of mass trees, chosen by the classifier. The usage of multitude of trees is to avoid overfitting.

The assessed performance of Random Forest is depicted by the confusion matrix below:

Table 2. Confusion Matrix of Random Forest

	TP	FP	TN	FN
DATASET 1	78	12	30	15
DATASET 2	1649	52	754	95

3.2.3 Naive Bayes Classifier:

Naive Bayes is a supervised learning algorithm in supposition which denotes all features as independent and unassociated with each other.

It works well with data with conditional problems.

Naive Bayes is a classification technique which implements the Bayes Theorem. Applying Bayes Theorem posterior probability P(A|B) can be computed from P(A),P(B) and P(B|A) .

Thence , P(A|B)=P(B|A)*P(A)/P(B)

Where,
P(A|B)=posterior probability
P(B|A)=likelihood
P(A)=prior probability
P(B)=evidence

The assessed performance of Naive Bayes classifier is depicted by the confusion matrix below:

Table 3. Confusion Matrix of Naive Bayes

	TP	FP	TN	FN
DATASET 1	75	14	28	14
DATASET 2	1562	139	515	334

3.2.4 Adaboost Classifier:

Adaboost is an unvarying ensemble technique. It incorporates diverse classifiers to extend their accuracy. The underlying conception responsible for Adaboost classifier is to line the weights of the classifiers and priming the sample in each subset specified iteratively as to stipulate the accurate predictions of unfamiliar observations.

It executes this by designating higher weights for erroneously classified observations to provide them with greater probability of being classified. This process repeats itself until the entire training data fits shorn of any error. The assessed performance of Adaboost classifier is depicted by the confusion matrix below:

Table 4. Confusion Matrix of Adaboost Classifier

	TP	FP	TN	FN
DATASET 1	84	7	34	13
DATASET 2	1731	50	833	71

3.3 Dataset Used:

This study was conducted on two datasets comprising of diabetic medical details.

PIDD-Pima Indians Diabetes Dataset
The aforementioned methodology is assessed on Diabetes Dataset viz. PIDD-Pima Indians Diabetes Dataset consisting of medical information of 768 instances which are female patients. The dataset incorporates 8 attributes in which value of one class '0' is used as tested negative for diabetes and value of an additional class '1' is processed as tested positive for diabetes.

Diabetes from DAT263xLab01
The antedecent emthodology is evaluated on Diabetes Dataset viz. Diabetes from DAT263xLab01 containing 15001 observations of females above the age of 21. It comprises of 9 attributes at which value of one class '0'is used as tested negative for diabetes and value of an additional class '1' is processed as tested positive for the disease.

Table 5. Attribute description

Attribute of PIMA India Diabetes Dataset (PIDD)	Attribute of dataset 2
Pregnancies	Patient Id
Glucose	Pregnancies
Blood Pressure	Glucose
Skin Thickness	Blood Pressure
Insulin	Skin Thickness
BMI	Insulin
Diabetes Pedigree Function	BMI
Age	Diabetes Pedigree Function
Class 0 or 1	Age
	Class 0 or 1

3.4 Accuracy Measures:

Decision tree, Adaboost, Random Forest and Naïve Bayes algorithm have been used in prediction of diabetes prediction. Efficiency of the algorithm were judged on the multi-dimensional basis i.e accuracy, F1, precision and recall. The

table below describes the attributes of the datasets used for carrying out this research.

Table 6. Accuracy Measures Description

Measuring Constraint	Significance	Formula
Accuracy	Accuracy determines the accuracy in prediction of an instance	A=(TP+TN) / (Total no of samples)
Precision	Classifier's Correctness/accuracy	P = TP / (TP+ FP)
Recall	Completeness and sensitivity of the classifier	R =TP / (TP+FN)
F1	Weighted avg of precision and recall	F=2*(P*R) / (P+R)

3.4.1 Comparative performance of different classification algorithms in accordance to accuracy measures.

Table 7. Accuracy Measures for the Datasets

Algorithm	Precision	Recall	F-Measure	Accuracy %
Random Forest	0.738	0.688	0.712	81.4
Adaboost	0.829	0.723	0.772	85.5
Decision Tree	0.666	0.65	0.658	79.4
Naïve Bayes	0.666	0.666	0.666	78.6

Algorithm	Precision	Recall	F-Measure	Accuracy %
Random Forest	0.937	0.893	0.915	94.4
Adaboost	0.943	0.921	0.932	95.4
Decision Tree	0.826	0.773	0.799	86.9
Naïve Bayes	0.706	0.393	0.505	74.3

The values of different classification algorithms and their performance and various criteria are listed in the above table.

TP signifies true positive, TN signifies true negative, FN signifies false negative and FP false positive.

Comparison of performance of the aforementioned algorithms are measured on the basis of precision, recall, F-measure and accuracy

4. RESULT:

It was inferred that Random forest proved to be more efficient than the other algorithms and achieved highest accuracy measures on PIDD dataset. However, Adaboost conferred better results on the second dataset.

The outcome of the experiments are summarized in the below graphs, where various accuracy measures are compared for Decision Tree, Random Forest, Adaboost and Naïve Bayes.

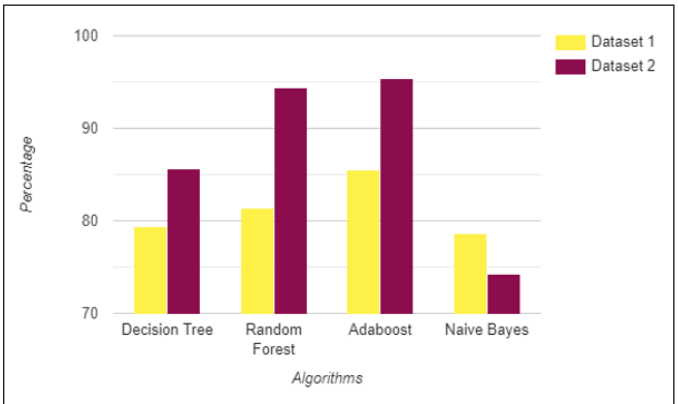

Fig. 2: Accuracy

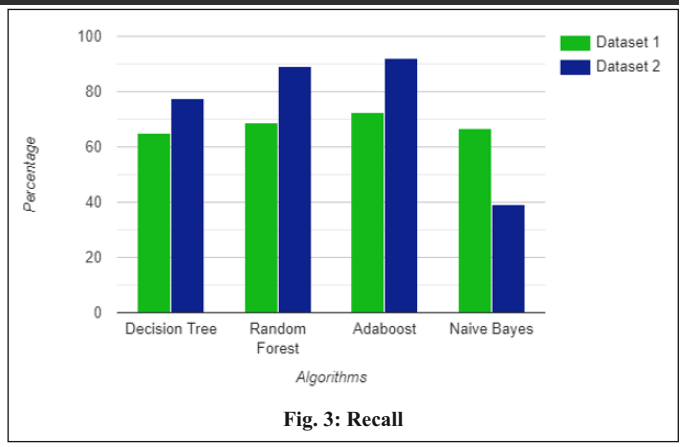

Fig. 3: Recall

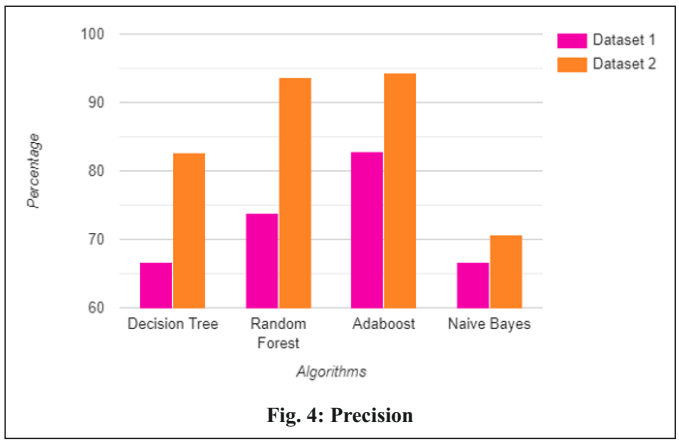

Fig. 4: Precision

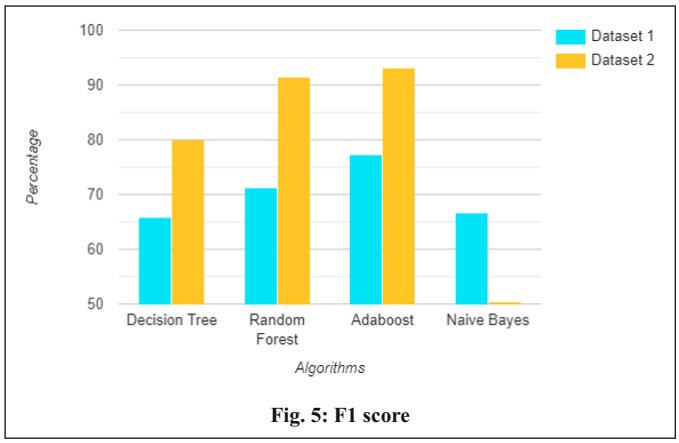

Fig. 5: F1 score

4. CONCLUSION:

Detection of Diabetes Mellitus at an early stage can help in taking precautionary measures. In this research various classification algorithms were put against each other and a systematic conclusion was drawn on the basis of performance of the algorithms on accuracy measures like precision, recall, F-Measure and accuracy. It was seen that for both datasets Adaboost classifier performed the best and has achieved comparatively better results. It achieved an accuracy of 85.5% and 95.4% for first and second datasets respectively.

REFERENCES:

I. Aishwarya, R., Gayathri, P., Jaisankar, N., (2013). A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908

II. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., (2013). Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

III. Arora, R., Suman, (2012). Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

IV. Bamnote, M.P., G.R., (2014). Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

V. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., (2017). Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the

International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.

VI. Dhomse Kanchan B., M.K.M., (2016). Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

VII. Esposito, F., Malerba, D., Semeraro, G., Kay, J., (1997). A comparative analysis of methods for pruning decision trees. IEEE Transactions on Pattern Analysis and Machine Intelligence 19, 476–491. doi:10.1109/34.589207.

VIII. Fatima, M., Pasha, M., (2017). Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

IX. Garner, S.R., (1995). Weka: The Waikato Environment for Knowledge Analysis, in: Proceedings of the New Zealand computer science research students conference, Citeseer. pp. 57–64.

X. Han, J., Rodriguez, J.C., Beheshti, M., (2008). Discovering decision tree based diabetes prediction model, in: International Conference onAdvanced Software Engineering and Its Applications, Springer. pp. 99–109.

XI. Iyer, A., S, J., Sumbaly, R., (2015). Diagnosis of Diabetes Using Classification Mining Techniques. International Journal of Data Mining & Knowledge Management Process 5, 1–14. doi:10.5121/ijdkp.2015.5101, arXiv:1502.03774.

XII. Iancu, I., Mota, M., and Iancu, E. (2008). "Method for the analysing of blood glucose dynamics in diabetes mellitus patients," in Proceedings of the 2008 IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca. doi: 10.1109/AQTR.2008.4588883

XIII. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., Chouvarda, I., (2017). Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal 15, 104–116. doi:10.1016/j.csbj.2016.12.005.

XIV. Krasteva, A., Panov, V., Krasteva, A., Kisselova, A., and Krastev, Z. (2011). Oral cavity and systemic diseases—Diabetes Mellitus. Biotechnol. Biotechnol. Equip. 25, 2183–2186. doi: 10.5504/BBEQ.2011.0022

XV. Kumar, D.A., Govindasamy, R., (2015). Performance and Evaluation of Classification Data Mining Techniques in Diabetes. International Journal of Computer Science and Information Technologies, 6, 1312–1319.

XVI. Kumar, P.S., Umatejaswi, V., (2017). Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications 7, 705–709.

XVII. Kumari, V.A., Chitra, R., (2013). Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications (IJERA) www.ijera.com 3, 1797–1801.

XVIII. Mujumdar, Aishwarya, and V. Vaidehi.(2019). "Diabetes prediction using machine learning algorithms." Procedia Computer Science 165 (2019): 292-299.

XIX. Nai-Arun, N., Moungmai, R.(2015). Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142. doi:10.1016/j.procs.2015.10.014.

XX. Nai-Arun, N., Sittidech, P.(2014). Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931 - 932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427.

XXI. Orabi, K.M., Kamal, Y.M., Rabah, T.M. (2016). Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.

XXII. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K.(2016). Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121. doi:10.1016/j.procs.2016.04.016.

XXIII. Pradhan, P.M.A., Bamnote, G.R., Tribhuvan, V., Jadhav, K., Chabukswar, V., Dhobale, V. (2012). A Genetic Programming Approach for Detection of Diabetes. International Journal Of Computational Engineering Research 2, 91–94.

XXIV. Priyam, A., Gupta, R., Rathee, A., Srivastava, S.(2013). Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and Technology Vol.3, 334–337. doi:JUNE 2013, arXiv:ISSN 2277 - 4106.

XXV. Ray, S.(2017). 6 Easy Steps to Learn Naive Bayes Algorithm (with code in Python).

XXVI. Rish, I.(2001). An empirical study of the naive Bayes classifier, in: IJCAI 2001 workshop on empirical methods in artificial intelligence, IBM. pp. 41–46.

XXVII. Robertson, G., Lehmann, E. D., Sandham, W., and Hamilton, D. (2011). Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. J. Electr. Comput. Eng. 2011:681786. doi: 10.1155/2011/681786

XXVIII. Sharief, A.A., Sheta, A.(2014). Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.

XXIX. Sisodia, D., Shrivastava, S.K., Jain, R.C.(2010). ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010, 554–559doi:10.1109/CICN.2010.109.

XXX. Sisodia, D., Singh, L., Sisodia, S.(2014). Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

XXXI. Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, (2016). An Intelligent Approach for Diabetes Classification, Prediction and Description. Advances in Intelligent Systems and Computing 424, 323–335. doi:10.1007/978-3-319-28031-8.

XXXII. Vijayan, V.V., Anjali, C.(2015). Prediction and diagnosis of diabetes mellitus A machine learning approach. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) , 122–127doi:10.1109/RAICS.2015.7488400.

XXXIII. Yu, W., Liu, T., Valdez, R., Gwinn, M., Khoury, M.J.(2010). Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making 10. doi:10.1186/1472-6947-10-16

XXXIV. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.